Language-driven All-in-one Adverse Weather Removal
Supplementary Material

We provide details of our LDR framework omitted in the main paper, and summarize this supplementary material into three sections: 1) Sec. 1, LDR implementation details; 2) Sec. 2, ablation studies; 3) Sec. 3, additional quantitative and qualitative comparisons.

1. Implementation Details

Our network architecture is given in Fig. 1. We adopt a multi-input and multi-output network architecture [4, 5, 13], incorporating the SFE block [4] and SF block [5] into our backbone network. Following each SF block, we append our language-driven restoration (LDR) block, which consists of degradation prior embedding, degradation map measurement, Top-K expert restoration, and restoration feature aggregation. We use a base channel of 32, and double/halve the number of channels after each downsampling/upsampling layer. Within our network, we utilize a pre-trained LISA-7B [6] to generate degradation priors.

The 'BestT' in Tab. 1 and Tab. 2 of the main paper denotes the best task-specific model. We report the best task-specific results by selecting state-of-the-art (SOTA) deraining [3], dehazing [15], desnowing [19] and raindrop removal [17] works. The best results are obtained by using state-of-the-art methods' pre-trained models directly, finetuning them, or retraining them.

2. Ablation studies and Discussions

We validate the effectiveness and components of our framework on the All-weather [10] dataset.

Pre-trained Vision-language Model. We compare pre-trained vision-language models (PVL) in generating degradation for our LDR framework, and the results are given in Tab. 1. Three state-of-the-art PVL models are studied, namely, the LISA [6], BLIP-2 [9], and miniGPT-v2 [1]. We achieve the best restoration performance with the degradation prior from LISA. Our LDR framework consistently outperforms the state-of-the-art method [20] with a PSNR/SSIM of 28.78 dB/0.909 (Refer to Table 1 in the main paper), regardless of the PVL model.

Table 1. *Comparison results of using different PVL models.*

PVL	PSNR$_{\uparrow}$	SSIM$_{\uparrow}$
BLIP-2	29.55	0.912
miniGPT-v2	29.63	0.914
LISA (Ours)	29.75	0.916

Table 2. *Comparison of degradation priors. NA is short for not applied.*

Degradation Type	W/. Reason	PSNR$_{\uparrow}$	SSIM$_{\uparrow}$
LISA	NA	29.43	0.910
LISA (Ours)	LISA	29.75	0.916
GT	NA	29.58	0.912
GT	LISA	29.84	0.917

Table 3. *Comparison of question prompts.*

Question Prompt	PSNR$_{\uparrow}$	SSIM$_{\uparrow}$
$\mathbf{T}_1$	29.66	0.915
$\mathbf{T}_2$	29.61	0.914
$\mathbf{T}_3$ (Ours)	29.75	0.916

Degradation Prior Component. Our degradation prior has two components, degradation type, and degradation reason (*e.g.*, severity and occurrence). The effectiveness of the two components is studied in Tab. 2. We compare the results of our model using three different degradation priors: 1) LISA classified degradation type, 2) ground truth (GT) degradation type, and 3) GT type augmented LISA, as the degradation prior separately.

For the GT type augmented LISA, the GT degradation type is provided to LISA to reason about degradation prior. With the inclusion of the degradation reason, the network is capable of improving the restoration performance. For example, GT type augmented LISA (29.84 dB) has a 0.26 dB improvement in PSNR compared to using GT Type alone (29.58 dB) as the degradation prior. Moreover, when compared with 'BestT + GT' that has 29.76 dB/0.916 in PSNR/SSIM from Tab. 1 of the main paper, GT type augmented LISA achieves a 0.08 dB higher PSNR. This shows the potential of an all-in-one framework to outperform task-specific models by leveraging the PVL model's capacity to learn accurate shared and task-specific knowledge.

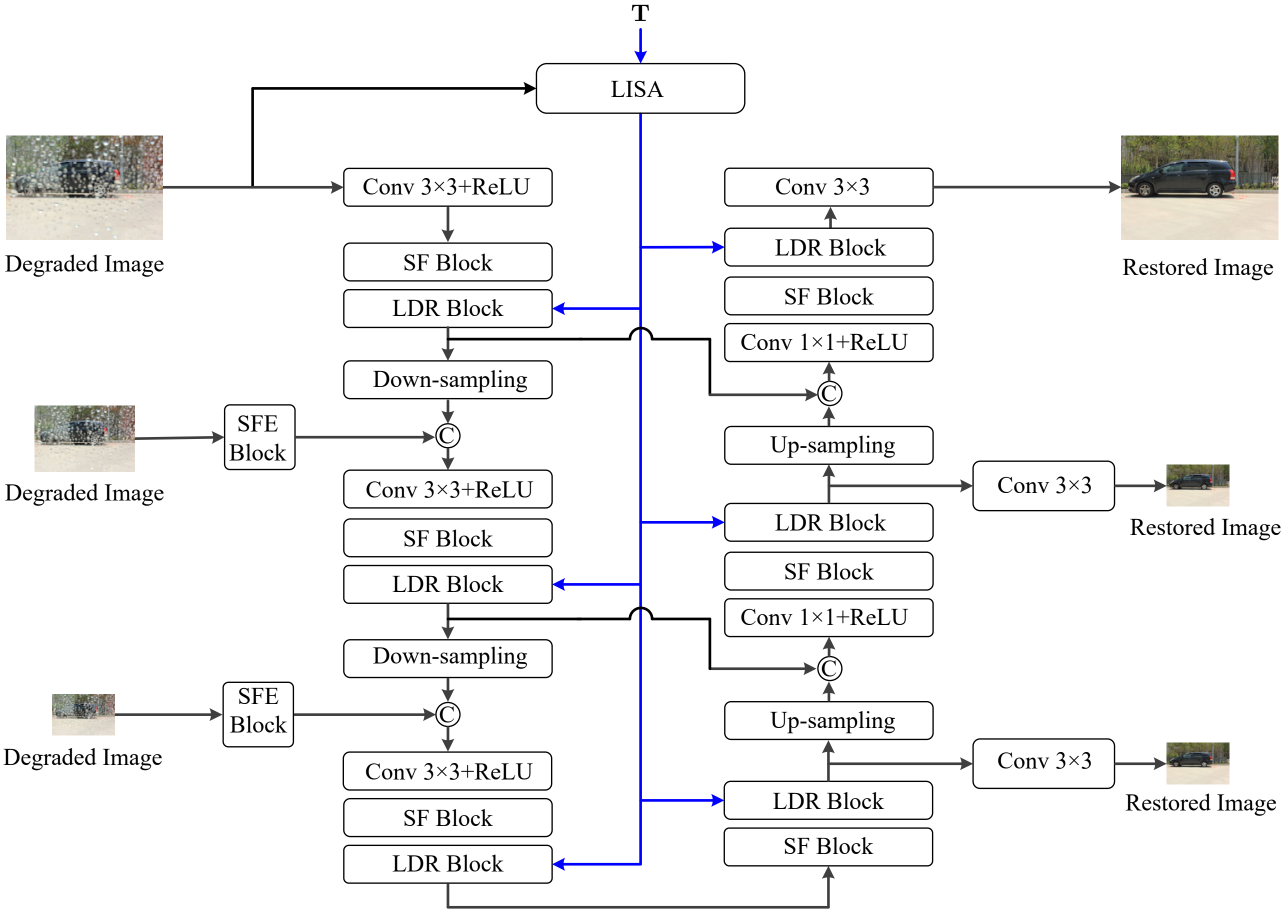


Figure 1. *LDR network architecture. We build a multi-input and multi-output network by using SFE [4], SF [5], and our LDR blocks. The LDR block is composed of our degradation prior embedding, degradation map measurement, Top-K expert restoration, and restoration feature aggregation modules as shown in our main paper. Here, '$\mathbf{T}$' is the question prompt, and '©' denotes concatenation. 'Conv 3×3' is a convolution layer with a kernel size of 3. 'Down-sampling' is a convolution layer with a stride of 2 and a kernel size of 3. 'Up-sampling' is a transposed convolution layer with a stride of 2 and a kernel size of 4.*

Question Prompt. We employ GPT-4, guided by the prompt [*Give me some sentences for asking a VQA model to describe the weather in a picture.*], to formulate the question prompt $\mathbf{T}$ for querying the degradation prior from LISA. The formatted prompts are 1) $\mathbf{T}_1$ = [*Can you analyze the weather conditions shown in this image?*], 2) $\mathbf{T}_2$ = [*Please describe about the weather in the picture.*], and 3) $\mathbf{T}_3$ = [*Please describe the type of weather, intensity, and obscured areas in the picture.*]. We provide examples of text descriptions obtained by using $\mathbf{T}_1$, $\mathbf{T}_2$, and $\mathbf{T}_3$ in Fig. 2. The comparison is given in Tab 3. The third prompt leverages LISA most for reasoning degradation prior, and achieves the highest performance.

Top-K Expert Restoration. We study the number of candidate experts N, selected top-K experts, and expert layers. The results are given in Tab. 4.

We start with a default configuration, *i.e.*, $N = 24$, top-K = 4, expert-layer = 4, and vary each of them independently. N is varied from 8 to 40 with a stride 8. Top-K and expert-layer are varied from 1 to 8 with a stride 2. Referring to Tab. 3 in the main paper as a reference baseline, without the Top-K Expert Restoration module ($N = 0$, top-K = 0, and expert-layer = 0), our model achieves 29.11 dB in PSNR and 0.902 in SSIM. We find that increasing the number of parameters improves the restoration performance. When $N = 24$, top-K = 4, and expert-layer = 4, the relative restoration performance improvement is decelerated, and the models with more parameters have similar performance. Thus, we use them for our framework for balancing the model efficiency and restoration performance.

Table 4. *Ablation study of the number of (a) candidate experts N, (b) selected top-K experts, and (c) expert layer. We vary one of the variables, fixing N = 24, top-K = 4, and expert-layer = 4. Two lightweight baseline settings are presented as references, 'B^1' is set to N = 2, Top-K = 1, and expert-layer = 1, and 'B^2' is set to N = 4, Top-K = 2, and expert-layer = 1.*

(a)

N	PSNR↑	SSIM↑
B^1	29.15	0.903
B^2	29.19	0.905
8	29.37	0.909
16	29.60	0.913
24	29.75	0.916
32	29.80	0.916
40	29.77	0.916

(b)

Top-K	PSNR↑	SSIM↑
B^1	29.15	0.903
B^2	29.19	0.905
1	29.50	0.910
2	29.61	0.913
4	29.75	0.916
6	29.78	0.916
8	29.80	0.916

(c)

expert-layer	PSNR↑	SSIM↑
B^1	29.15	0.903
B^2	29.19	0.905
1	29.23	0.907
2	29.53	0.912
4	29.75	0.916
6	29.78	0.916
8	29.83	0.917

Table 5. *Quantitative comparison on the restoration of images with mixed degradation.*

Method	**Snow and Haze**		**Raindrop and Snow**		**Haze and Raindrop**		**Average**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
AirNet [8]	18.66	0.632	15.22	0.705	16.12	0.752	16.67	0.696
WGWS [20]	20.12	0.667	17.32	0.735	19.23	0.772	18.89	0.724
Ours	22.02	0.705	21.32	0.850	22.14	0.873	21.82	0.809

Y-channel Evaluation. Considering several existing approaches use Y-channel PSNR and SSIM as the evaluation metrics, we report the values of our method on the All-weather dataset as 32.27 (dB) and 0.935, respectively.

Computational costs. Our method is trained on RTX A6000s, and the parameter numbers, inference time, and FLOPs are given below.

Method	Param.	Infer.	FLOPs
AirNet [8]	8.9M	0.13s	290G
WGWS [20]	24.4M	0.15s	212G
WDiff [14]	86M	17.16s	13425G
PVL + Ours	7B + 14M	0.39s + 0.08s	635G + 97G

Future Work and Limitation. Our model can be potentially applied to broader image restoration tasks, and strengthen by using a stronger PVL model than LISA, *e.g.*, the recent availability of GPT-4Vision since November 6th, 2023, presents an opportunity for strengthening our model.

In our model, 83% of the inference time is attributed to LISA. The model inference speed can be improved by distilling a smaller PVL model than LISA that is customized to adverse weather. If the intensity of rain is high, it creates splattering effects when it hits the surface of objects in the scene. Eliminating this effect remains a challenge for all methods, including ours.

3. Additional Results

Mixed Degradation. Adverse weather degradation is usually mixed by multiple degradation types [18]. Hence, we evaluate the generalization ability of our model on restoring images degraded by mixed/unencountered weather in real-world applications. As shown in Fig. 3, our model and SOTA methods (trained on the All-weather dataset [10]) are tested on real images with mixed and unencountered degradation, *e.g.*, 'snow and rain' (1^{st} row), 'snow and haze' (2^{ed}-3^{rd} row), 'snow, rain, and raindrop' (4^{th} row), as well as 'rain and haze' (5^{th}-6^{th} row). Moreover, we provide quantitatively compare with past methods. We randomly select 300 single-degradation images from the All-weather dataset, and follow Liu *et al.* [12] and Li *et al.* [7] to synthesize images of mixed degradation. The comparison results are shown in Tab. 5 and Fig. 3. Compared to AirNet [8] and WGWS [20], our method achieves a gain of 5.15 dB/0.107 and 2.93 dB/0.085 in PSNR/SSIM with cleaner restorations, showing a better generalization ability on restoration of images with mixed/unencountered degradation.

All-in-one Restoration. We provide more quantitative comparison results on the All-weather dataset and the WeatherStream dataset. The compared baselines are GRL [11], AirNet [8], TUM [2], Transweather [16], and WGWS [20]. The results are shown in Fig. 4, Fig. 5, and Fig. 6.

References

[1] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1

[2] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17653–17662, 2022. 3, 7, 8, 9

[3] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5896–5905, 2023. 1

[4] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 1, 2

[5] Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, and Alois Knoll. Selective frequency network for image restoration. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[6] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1, 5

[7] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 3

[8] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17452–17462, 2022. 3, 6, 7, 8, 9

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[10] Ruoteng Li, Robby T Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3175–3185, 2020. 1, 3

[11] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023. 3, 7, 8, 9

[12] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018. 3

[13] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1905–1913, 2023. 1

[14] Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[15] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 1

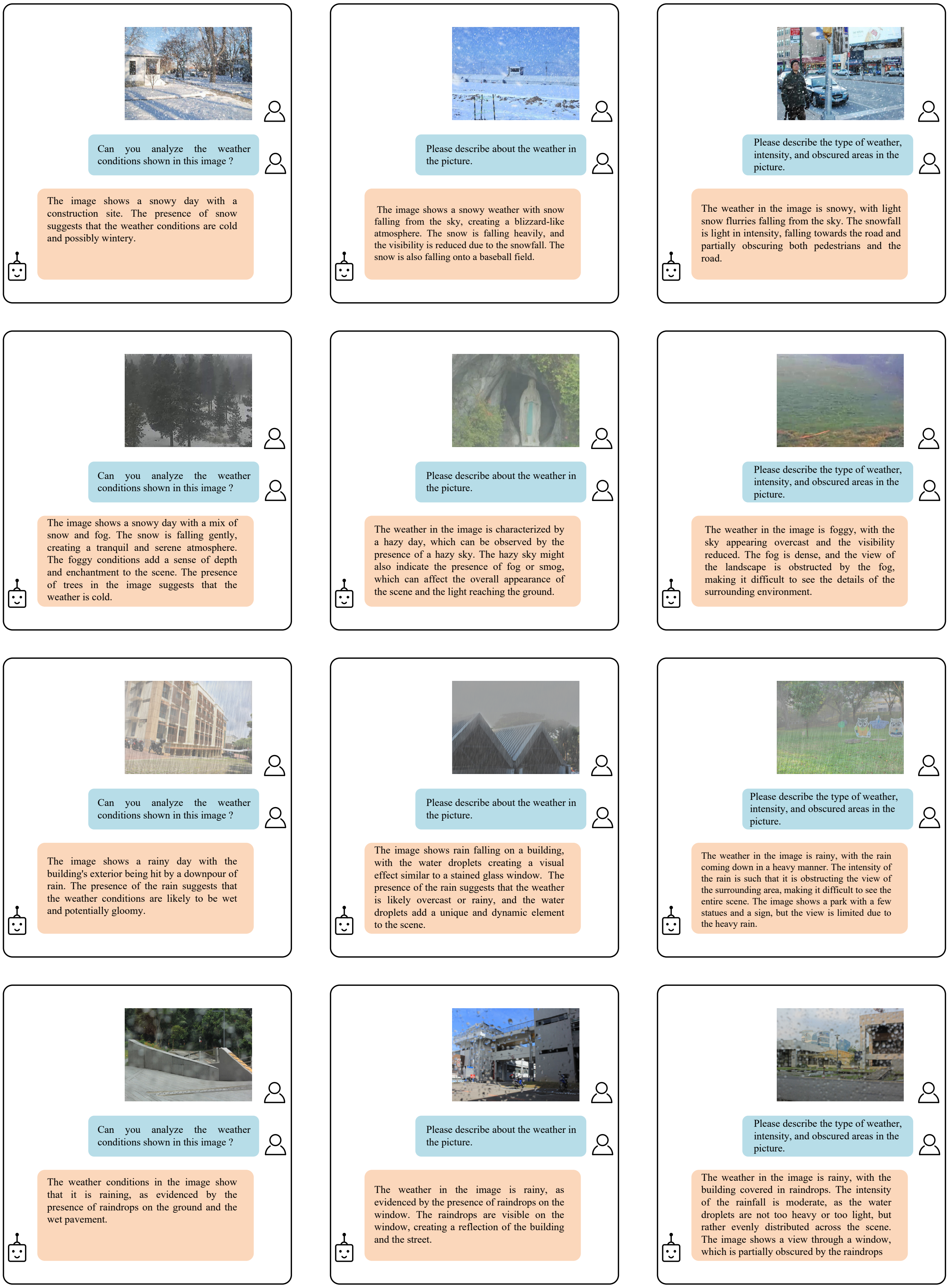
[16] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2363, 2022. 3, 7, 8, 9

[17] Yinglong Wang, Chao Ma, and Jianzhuang Liu. Smartassign: Learning a smart knowledge assignment strategy for deraining and desnowing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2023. 1

[18] Kaihao Zhang, Dongxu Li, Wenhan Luo, and Wenqi Ren. Dual attention-in-attention model for joint rain streak and raindrop removal. *IEEE Transactions on Image Processing*, 30:7608–7619, 2021. 3

[19] Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. Deep dense multi-scale network for snow removal using semantic and depth priors. *IEEE Transactions on Image Processing*, 30:7419–7431, 2021. 1

[20] Yurui Zhu, Tianyu Wang, Xueyang Fu, Xuanyu Yang, Xin Guo, Jifeng Dai, Yu Qiao, and Xiaowei Hu. Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21747–21758, 2023. 1, 3, 6, 7, 8, 9



Can you analyze the weather conditions shown in this image ?

The image shows a snowy day with a construction site. The presence of snow suggests that the weather conditions are cold and possibly wintery.

Please describe about the weather in the picture.

The image shows a snowy weather with snow falling from the sky, creating a blizzard-like atmosphere. The snow is falling heavily, and the visibility is reduced due to the snowfall. The snow is also falling onto a baseball field.

Please describe the type of weather, intensity, and obscured areas in the picture.

The weather in the image is snowy, with light snow flurries falling from the sky. The snowfall is light in intensity, falling towards the road and partially obscuring both pedestrians and the road.

Can you analyze the weather conditions shown in this image ?

The image shows a snowy day with a mix of snow and fog. The snow is falling gently, creating a tranquil and serene atmosphere. The foggy conditions add a sense of depth and enchantment to the scene. The presence of trees in the image suggests that the weather is cold.

Please describe about the weather in the picture.

The weather in the image is characterized by a hazy day, which can be observed by the presence of a hazy sky. The hazy sky might also indicate the presence of fog or smog, which can affect the overall appearance of the scene and the light reaching the ground.

Please describe the type of weather, intensity, and obscured areas in the picture.

The weather in the image is foggy, with the sky appearing overcast and the visibility reduced. The fog is dense, and the view of the landscape is obstructed by the fog, making it difficult to see the details of the surrounding environment.

Can you analyze the weather conditions shown in this image ?

The image shows a rainy day with the building's exterior being hit by a downpour of rain. The presence of the rain suggests that the weather conditions are likely to be wet and potentially gloomy.

Please describe about the weather in the picture.

The image shows rain falling on a building, with the water droplets creating a visual effect similar to a stained glass window. The presence of the rain suggests that the weather is likely overcast or rainy, and the water droplets add a unique and dynamic element to the scene.

Please describe the type of weather, intensity, and obscured areas in the picture.

The weather in the image is rainy, with the rain coming down in a heavy manner. The intensity of the rain is such that it is obstructing the view of the surrounding area, making it difficult to see the entire scene. The image shows a park with a few statues and a sign, but the view is limited due to the heavy rain.

Can you analyze the weather conditions shown in this image ?

The weather conditions in the image show that it is raining, as evidenced by the presence of raindrops on the ground and the wet pavement.

Please describe about the weather in the picture.

The weather in the image is rainy, as evidenced by the presence of raindrops on the window. The raindrops are visible on the window, creating a reflection of the building and the street.

Please describe the type of weather, intensity, and obscured areas in the picture.

The weather in the image is rainy, with the building covered in raindrops. The intensity of the rainfall is moderate, as the water droplets are not too heavy or too light, but rather evenly distributed across the scene. The image shows a view through a window, which is partially obscured by the raindrops

Figure 2. *Examples of text descriptions from LISA [6] for reasoning weather degraded images. From left to right, descriptions are generated using* $\mathbf{T}_1$, $\mathbf{T}_2$, *and* $\mathbf{T}_3$, *respectively. From top to bottom, the degradation is snow, haze, rain and raindrop.*

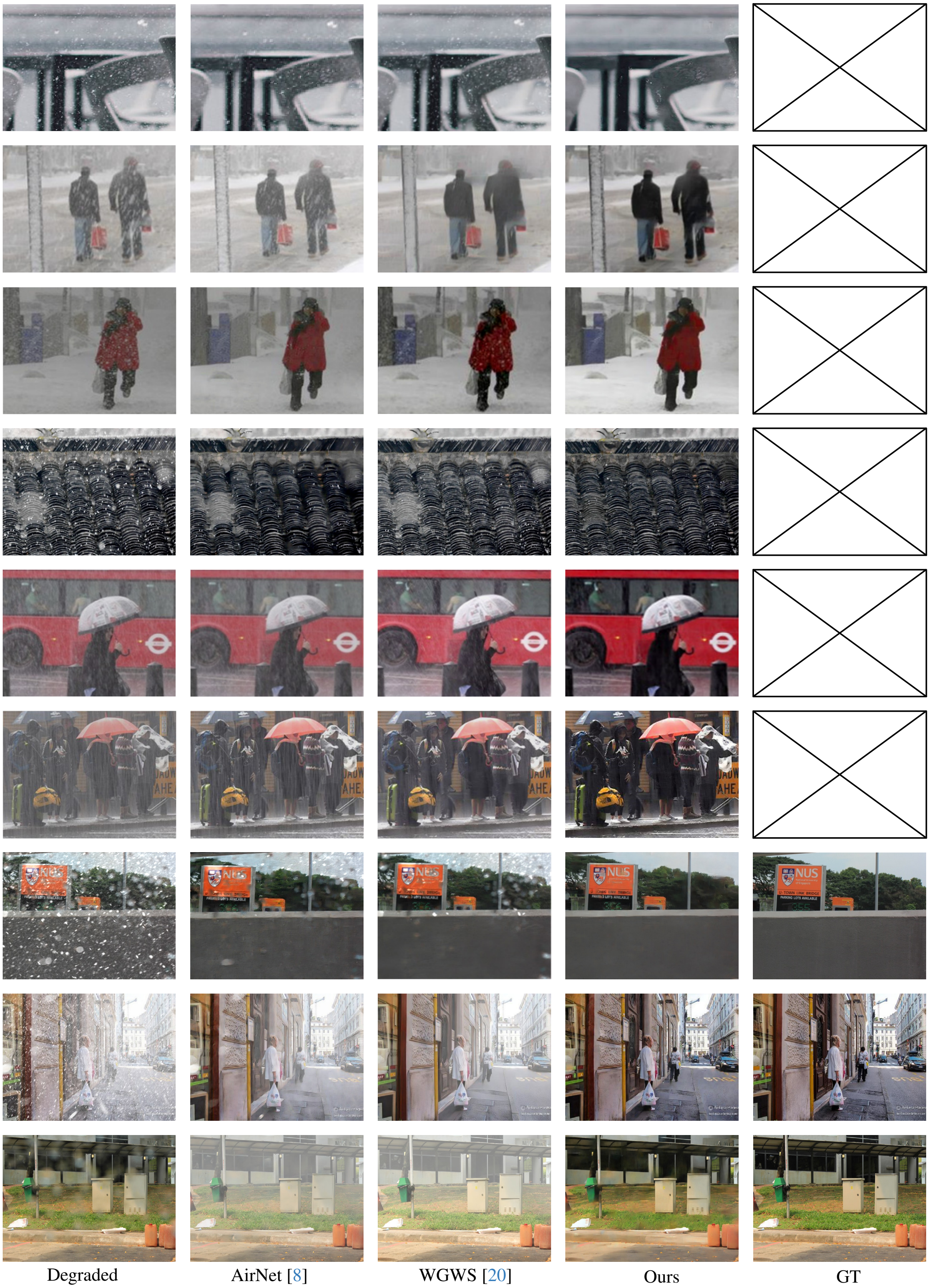


Figure 3. *Quantitative evaluation on our synthesized mixed weather dataset. The top six rows are restorations for mixed degradation in the real scene, and the bottom three rows are restorations for the synthesized mixed degradation.*

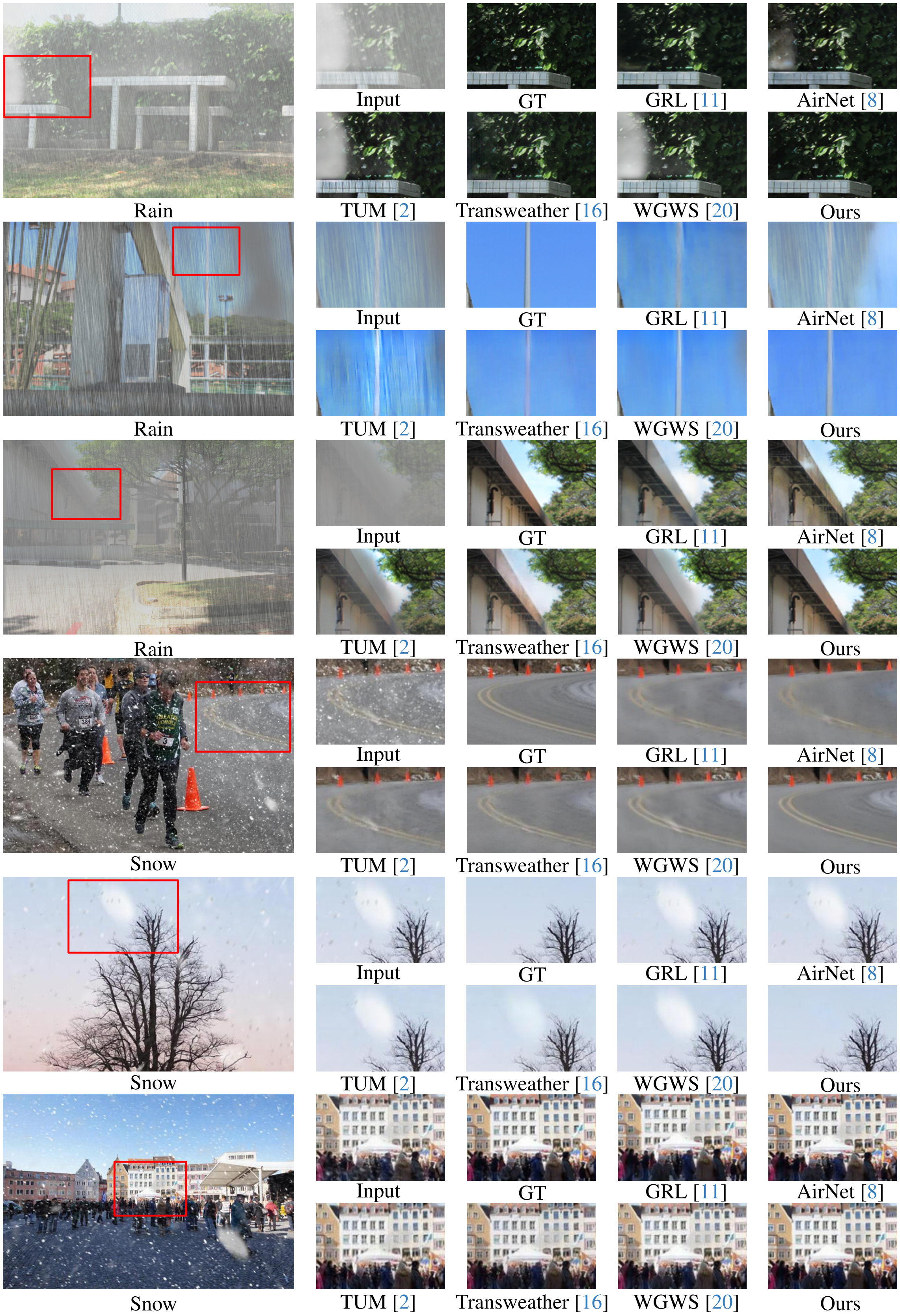


Figure 4. *Qualitative comparison on the All-weather dataset. The first column shows degraded images, while the crops for the bounding box regions of degraded images, ground truth, restorations from SOTA methods and our method are shown in the subsequent columns.*

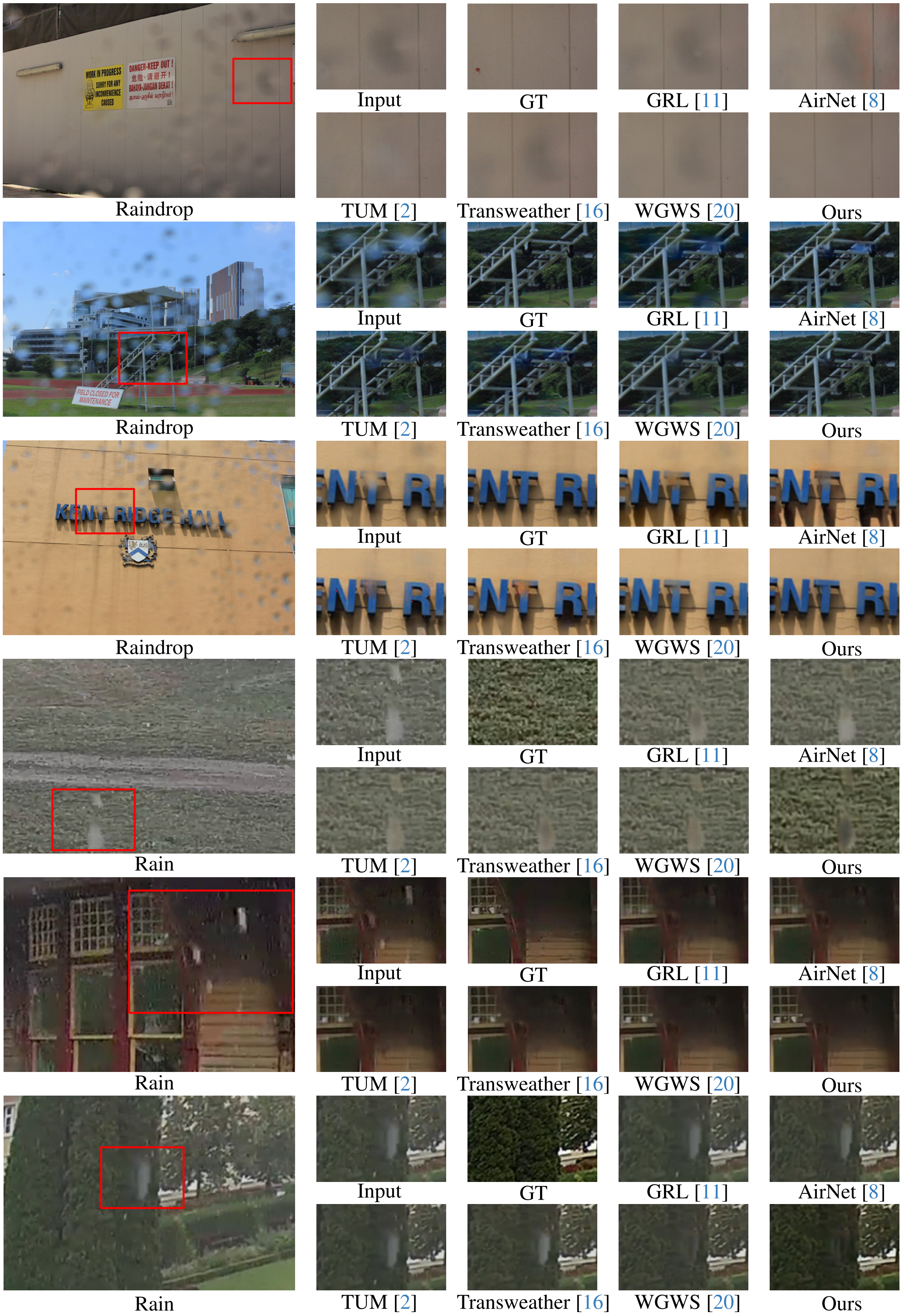


Figure 5. *Qualitative comparison on the All-weather (the first three rows) and WeatherStream (the last three rows) dataset. The first column shows degraded images, while the crops for the bounding box regions of degraded images, ground truth, restorations from SOTA methods and our method are shown in the subsequent columns.*

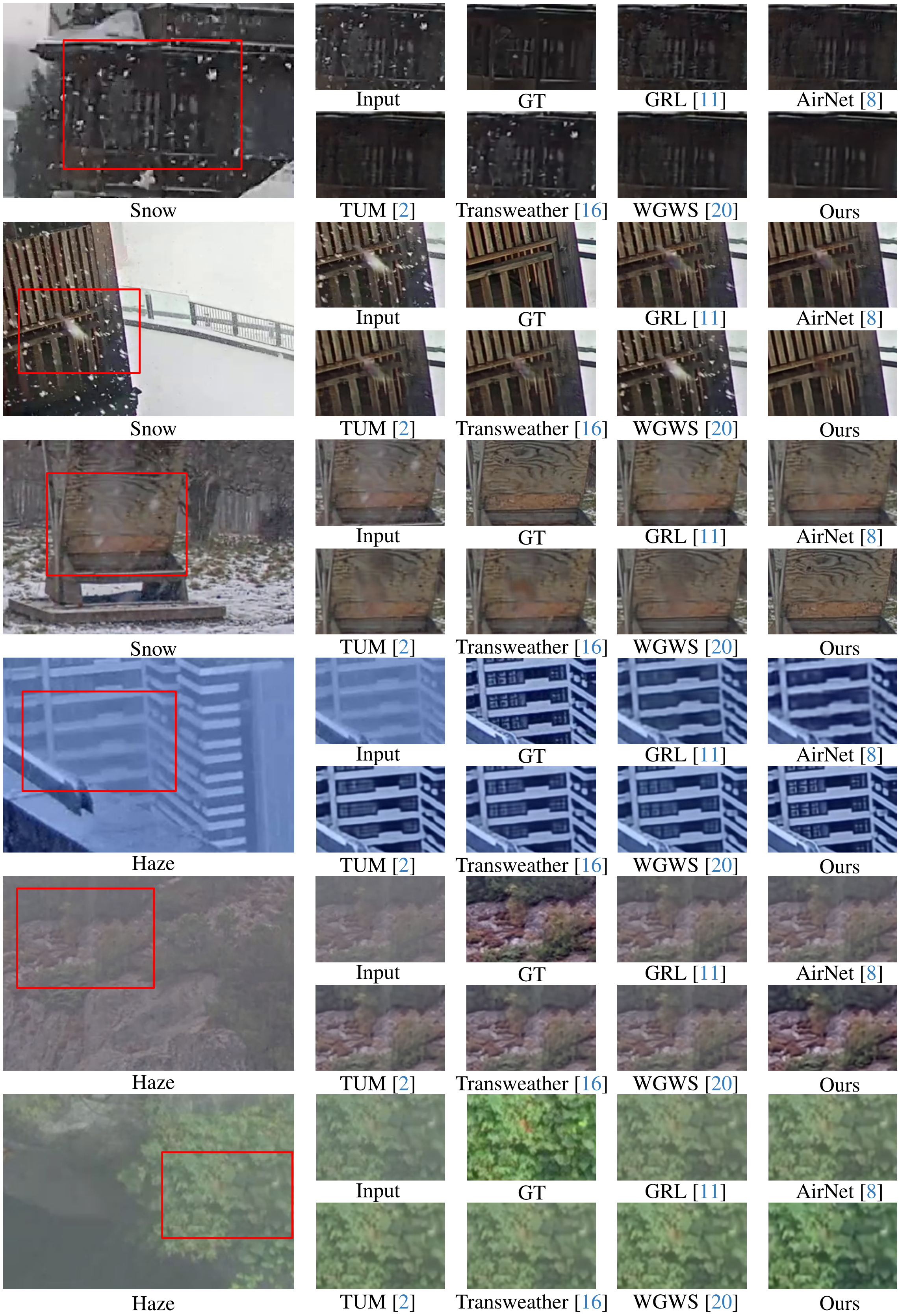


Figure 6. *Qualitative comparison on the WeatherStream dataset. The first column shows degraded images, while the crops for the bounding box regions of degraded images, ground truth, restorations from SOTA methods and our method are shown in the subsequent columns.*